International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 6, Issue 1, Feb 2016, 79-86
© TJPRC Pvt. Ltd.

TRANS STELLAR
•Journal Publications •Research Consultancy

Original Article

SECURE AND EFFICIENT MULTI-KEYWORD RANKED SEARCH OVER ENCRYPTED CLOUD DATA WITH RECENTLY UPDATED RESULTS

PREETHI MATHEW

Research Scholar, Department of Computer Science Engineering, SNGCE, Kadayirippu, Kerala, India

ABSTRACT

As the popularity of cloud computing is increasing, more and more sensitive information is being outsourced into the cloud. Sensitive data have to be encrypted before outsourcing to preserve data privacy. This usually implies that one has to sacrifice effective data utilization for security. Moreover, large number of documents and data users in cloud demand usage of multiple keyword in search request. Here, a solution is formulated for the problem of effective and secure utilization of cloud data which yields most relevant documents. A cloud data hosting service with three different entities are considered here namely the data owner, data user and cloud server. The data owner encrypts the collection of documents using AES algorithm. The data owner also builds an encrypted searchable index from the collection of documents. Stemming technology and homomorphic encryption is employed here. Both the index and encrypted document collection are outsourced to the cloud server. An authorized user gets a trapdoor to search the document collection for the given keywords. Upon receiving trapdoor from the user, the cloud server returns relevant documents after ranking the search result. The efficient similarity measure of "coordinate matching" is the ranking criteria used. The user can send a number k and the server sends back only top-k documents that are most relevant to the search query. The user can make use of the documents by downloading and decrypting it. Provisions are also made to update the data collection by inserting new documents. The proposed scheme impose low overhead on computation and communication.

KEYWORDS: Cloud Computing, Ranked Search, Encryption, Index

Received: Dec 17, 2015; **Accepted:** Jan 19, 2016; **Published:** Jan 27, 2016; **Paper Id.:** IJCSEITRFEB20169

INTRODUCTION

Cloud computing has widely emerged as a technology for data outsourcing and high quality data services in the IT sector as it is flexible, faster and cheaper. Cloud computing is delivery of computing as a service where we can globally and remotely store huge amount of data. Cloud platforms like Microsoft Azure, Amazon S3, Drop Box, Google Drive, Sky Drive provides storage services**.**

Data owners and enterprises are motivated to outsource their complex data management systems to cloud due to economic savings. Privacy problems may arise as sensitive information’s like personal records, tax and financial documents, photos are uploaded to cloud. In order to avoid unauthorized access to these sensitive data, the data have to be encrypted before outsourcing to cloud but efficiency will be compromised.

Keyword-based retrieval allows users to retrieve files they are interested in and it is prominently used in plaintext search schemes. Numerous searchable symmetric encryption schemes have been proposed to enable search on cipher text but they support only Boolean keyword search. Here we focus on Searchable Symmetric Encryption (SSE).This enables user to search encrypted data as documents and search securely through multi-keyword. As there are large amount of documents in cloud there is a need for ranking the files in the order of relevance by user’s interest. It helps users to retrieve relevant information quickly and eliminates unnecessary network traffic. With the usage of multiple keywords in the search request, each keyword will help to reduce the search result.

In this paper we define and solve the problem of secure and efficient multi-keyword ranked search over encrypted cloud data with recently updated results (SEMRSE).The cloud server is curious to analyze the data in its storage to learn additional information. This is considered as a threat for cloud security. Keeping this in mind the system design focus on simultaneously achieving security and performance Vector space model is used for index construction where each document is associated with a vector. Here tf-idf values are round to integers and then it is encrypted using homomorphic encryption to protect the index privacy. Stemming technology is used for constructing the keyword list as it improves efficiency. The search query is associated with binary vector and each bit represents whether the keyword is present in the search request. Dimension extending is incorporated in both data vector and query vector by assigning random numbers to it. The randomness will increase the difficulty for the cloud server to learn additional information from the data stored and message flows in the cloud. The inner product similarity is used to evaluate similarity measure of document to the search query. The user can update the document collection by inserting new documents and the user can retrieve the recently uploaded documents relevant to their query.

RELATED WORK

Existing searchable encryption schemes allows a user to securely search over encrypted data through keywords without decrypting it. Goh [8] proposed the construction of indexes using Bloom filters for every data files. Traditional single keyword searchable encryption schemes [6], [8], [16], [17], [18], [19] allow a user to search with single keyword. Here, an encrypted searchable index is built and its content is hidden from the server. Single keyword search without ranking will provide irrelevant results. In [10], [20] keyword frequency is used to rank results but here multi keyword search is not possible. All existing schemes support exact keyword search which is unsuitable for cloud computing.

Boolean keyword search schemes [4], [7], [9], [11], [12], [14] are still not adequate to provide users with acceptable result ranking functionality. Boolean keyword search provides true or false scenario without relevance ranking. Conjunctive keyword search only returns those documents where all the keywords in the search query appear. Disjunctive keyword search returns only that document that contains a subset of the specific keywords in the search query. Encryption schemes [11], [12], [14] recently proposed support both conjunctive and disjunctive search. None of existing Boolean keyword searchable encryption schemes support multiple keywords ranked search over encrypted cloud data while preserving privacy.

Ranking based search can eliminate unnecessary network traffic as it returns only relevant results. To improve user searching experience it is necessary to support multi keyword search to retrieve most relevant data as each keyword in the search request help to reduce the search result.

PROBLEM FORMULATION

System Model

A cloud computing system hosting data service is considered here in which three different entities are present, data owner, data user and cloud server. The data can contain much sensitive information. The data owner has a collection of documents F and should outsource F in the encrypted form C, as the cloud servers cannot be completely trusted to protect data. The cloud server will provide keyword retrieval service to authorized users. The data owner builds a searchable index I from F and then outsources the encrypted index and the encrypted files C onto the cloud server. The data user at first generates a query and the keywords are kept concealed for privacy reasons. An authorized user acquires a trapdoor T to search the document collection with the given keywords. Corresponding set of encrypted documents is returned to the user upon receiving T from data user. An number k can be send along with the trapdoor T by the user to reduce the network traffic, as it sends back only top-k documents that are recently updated and most relevant to the query. The data user can use the files after decrypting it.

Design Goals

In this paper, we address the problem of secure and efficient multi keyword ranked search over encrypted cloud data. For effective utilization of outsourced cloud data, we have the following goals.1) to design a mechanism for constructing keyword sets which are storage efficient.2) to design a multi keyword search scheme based on the constructed keyword sets.3) to provide result similarity ranking for effective data retrieval.4) to meet privacy requirements the cloud server is prevented from learning additional information from data set and index.5) to provide low computation and communication overhead.6) to enable users to upload new documents to the cloud. 7) To enable users to retrieve recently uploaded results.

SEMRSE FRAMEWORK

In our design, first we construct storage efficient keyword sets with Stemming technology. Then an efficient privacy preserving search scheme is proposed. The dimension extending operation is preserved and in addition a new random number t is assigned to the extended dimension in each query vector, which makes it difficult for the cloud server to learn the relationship among the trapdoors. Also, a random keyword is inserted to each data vector. Each individual vector is extended to (n+2) dimension. The whole scheme to achieve secure and efficient ranked search with multiple keywords over encrypted data with recently updated results is as follows:

- **Setup (λ):** The data owner generates the secret key and public keys for the homomorphic encryption scheme. The security parameter λ is taken as the input; the output is a secret key S, and a public key set P.

- **Index Build(C, P):** The data owner builds the secure searchable index from the file collection C .The data owner extracts the collection of l keywords, W = {w1, w2, ..., wl}, and their TF and IDF values out of the collection of n files, C = {f1; f2; . . . ; fn}.Stemming algorithm is used here for reducing inflected words to their root words, which improves efficiency. For each file fi Є C, the data owner builds a l dimensional vector vi={$t_{1,j}$,$t_{2,j}$,….$t_{i,l}$} where each term in vi represents tf-$idf_{wj,fi}$ (1≤ j ≤ l). The searchable index I= {v_i|1≤ i≤ n}. The data owner encrypts the searchable index I to secure searchable index I $^{‘}$= {$v_i^{’}$|1 ≤ i ≤ n} using homomorphic encryption. Each vector $v_i^{’}$ is then converted to (l+2) dimension vector where (l+1)th entry is set to a random number and (l+2)th entry is set to 1. The data owner encrypts C into C$^{’}$ = {f1$^{’}$, f2$^{’}$, f3’……….fn$^{’}$} with AES algorithm. And then outsources C$^{’}$ and I $^{‘to}$ the cloud server.

- ***Trapdoor Gen(R):*** A multi keyword query R is generated by the user. Here m keywords in R is given as input .A binary vector T is generated where each of binary bit T[j] represents whether the keyword is present in the keyword list W. B is extended to (l +2) dimension vector T$^{’}$. In T$^{’}$ the (l +1) th bit is set to 1 then it’s scaled by a random number r where r ≠ 0, after which it’s extended to (l +2) dimension with another random number b. Hence T’ is the trapdoor generated.

- ***Score Calculate*** ***(T', I'):*** When cloud server receives secure trapdoor T', for each file vector v_i' in I', the cloud server computes the inner product $p_i' = v_i'.T'$ $(1 \leq i \leq n)$ and returns the encrypted result vector E back to the data user. Here $E = \{v_1' p_1' \ldots v_i' p_i'\}$ where $(1 \leq i \leq n)$.

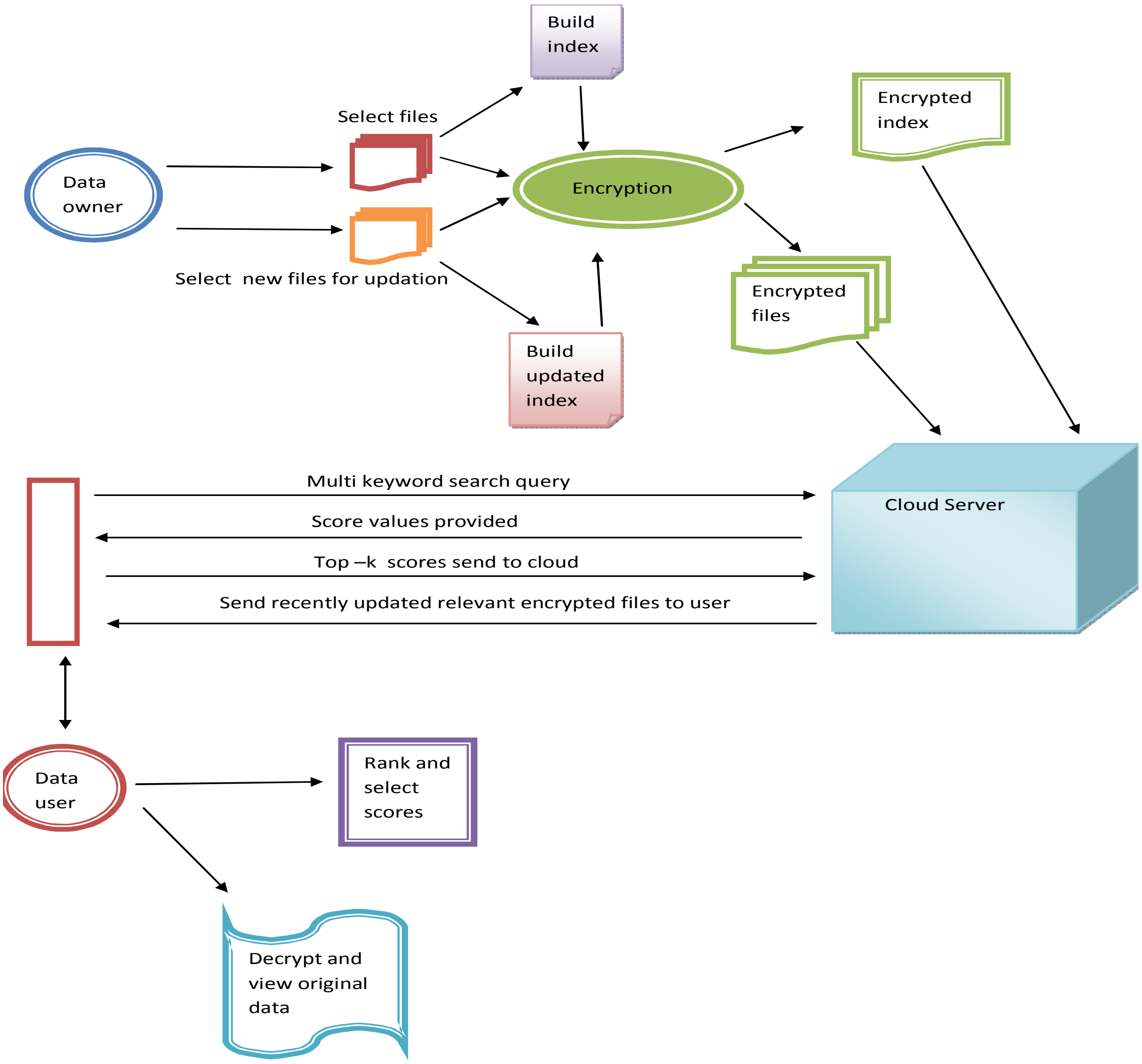
Build index
Encrypted index
Select files
Data owner
Encryption
Select new files for updation
Encrypted files
Build updated index
Multi keyword search query
Cloud Server
Score values provided
Top –k scores send to cloud
Send recently updated relevant encrypted files to user
Data user
Rank and select scores
Decrypt and view original data

Figure 1: SEMRSE Design

- ***Rank (E, S, and k)*****:** The data user decrypts the result vector E with secret key S and then requests and gets the files with top-k scores, where k is specified by the user. The cloud server returns the encrypted k files to the data user.

- ***Update (C):*** The owner can upload new documents to the cloud server by inserting new keywords to the blank spaces in the dictionary. Generate sub indexes for new documents based on the updated dictionary. The other documents and their sub indexes stored on the cloud server are not affected and therefore remain the same as before. The term frequency of newly uploaded documents is combined with an additional score so that the user retrieves recently uploaded documents relevant to their query.

SYSTEM ANALYSIS

An experimental evaluation of the existing system and the proposed system is performed on real world documents. The system is implemented using Java language with Intel Core i5 Processor, 4 GB RAM, 2.30GHz.

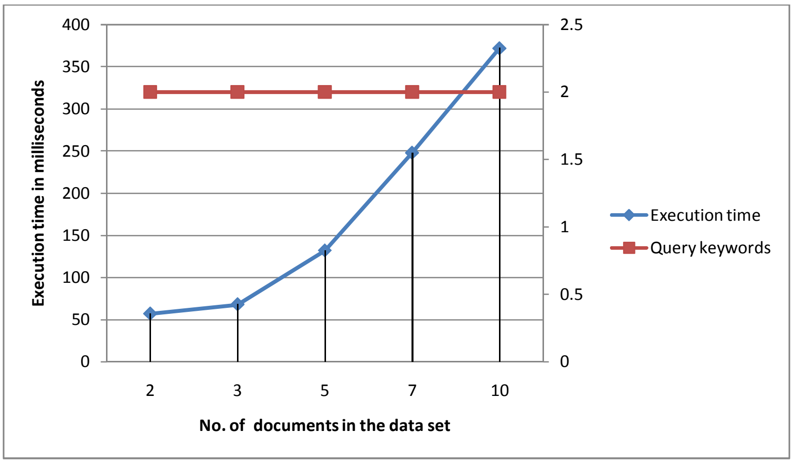

Figure 2: Execution Time Graph for different Number of Documents in Data set in SEMRSE

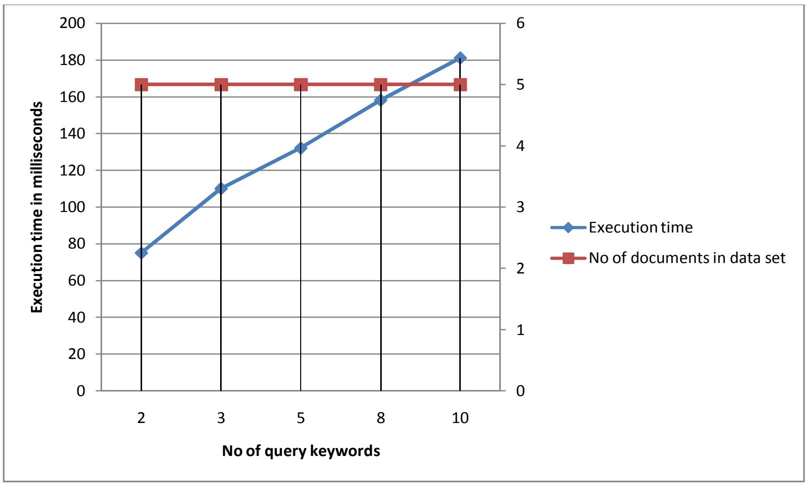

Figure 3: Execution Time Graph for different Number of Query Keywords in SEMRSE

As seen in Figure 2 randomly select different number of documents and form different data sets. Then a query is given to each data set and the execution time for score calculation is noted for each data set. Here, the number of query keywords is kept constant as 2. The effect of query keywords on the execution time is shown in Figure 3 by keeping the number of documents in the data set constant. Here the number of documents in the data set is kept as 5.Then the execution time for score calculation is determined for different number of query keywords. From both the graph it is clear that the execution time increases as number of documents in the data set and/or the number of keywords in the query increases.

CONCLUSIONS

The proposed method SEMRSE, motivate and solve the problem of secure multikeyword ranked retrieval over encrypted cloud data. The scheme employs fully homomorphic encryption which fulfills the security requirements of multikeyword ranked retrieval over the encrypted cloud data. Majority of computing work is done on the server side by operations only on cipher text. Stemming technology is used for reducing inflected words to their root words and it reduces the size of keyword list and improves efficiency of the system. The vector space model provides search accuracy. This scheme allows search request with multiple keywords and return documents in the order of their relevance to these keywords. Ranked search eliminates unnecessary network traffic by sending back only the most relevant data, which is suitable in the cloud paradigm. Also the user retrieves recently updated results from the cloud server, relevant to the search query. Among various multi-keyword semantics, the efficient similarity measure of "coordinate matching," i.e., as many matches as possible is used, to effectively capture the relevance of outsourced documents to the query keywords, and use "inner product similarity" to quantitatively evaluate such similarity measure. The proposed scheme guarantees data privacy. According to the efficiency evaluation of the proposed scheme over a real data set, extensive experimental results demonstrate that the proposed scheme has low computational overhead than the existing schemes and ensures practical efficiency.

FUTURE WORK

In the proposed system we have considered only text documents .In future we can extend this concept for images and videos too. As a future work, we can explore checking the integrity of the rank order in the search result assuming the cloud server is untrusted. Checking the integrity of ranked order is important because the order sent by cloud and the order received by data user may be changed by the attack of the intruder. Provisions can also be made to incorporate deletion of uploaded documents by the owner from the cloud server.

REFERENCES

1. *N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc. IEEE INFOCOM, pp. 829-837, Apr, 2014.*
2. *Jiadi Yu, Peng Lu, Yanmin Zhu, Guangtao Xue, Member, IEEE Computer Society, and Minglu Li, "Toward Secure Multikeyword Top k Retrieval over Encrypted Cloud Data", IEEE Transactions, July 2013.*
3. *Ming Li et al.," Authorized Private Keyword Search over Encrypted Data in Cloud Computing, IEEE proc. International conference on distributed computing systems, June 2011, pages 383-392.*
4. *D. Boneh and B. Waters, "Conjunctive, Subset, and Range Queries on Encrypted Data," Proc. Fourth Conf. Theory Cryptography (TCC), pp. 535-554, 2007.*
5. *Ming Li et al.,"Toward Privacy-Assured and Searchable Cloud Data Storage Services", IEEE Transactions on Network, volume 27, Issue 4, July/August 2013.*
6. *D. Boneh, G.D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public Key Encryption with Keyword Search," Proc. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2004.*
7. *R. Brinkman, "Searching in Encrypted Data," PhD thesis, Univ. of Twente, 2007.*
8. *E.-J. Goh, "Secure Indexes," Cryptology e Print Archive, http://eprint.iacr.org/2003/216. 2003.*

9. Y. Hwang and P. Lee, "Public Key Encryption with Conjunctive Keyword Search and Its Extension to a Multi-User System,"Pairing, vol. 4575, pp. 2-22, 2007.

10. C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure Ranked Keyword Search over Encrypted Cloud Data," Proc. IEEE 30th Int'l Conf. Distributed Computing Systems (ICDCS'10), 2010.

11. J. Katz, A. Sahai, and B. Waters, "Predicate Encryption Supporting Disjunctions, Polynomial Equations, and Inner Products," Proc. 27th Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2008.

12. A. Lewko, T. Okamoto, A. Sahai, K. Takashima, and B. Waters, "Fully Secure Functional Encryption: Attribute-Based Encryption and (Hierarchical) Inner Product Encryption," Proc. 29th Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT '10), 2010.

13. A. Swaminathan, Y. Mao, G.-M. Su, H. Gou, A.L. Varna, S. He, M. Wu, and D.W. Oard, "Confidentiality-Preserving Rank-Ordered Search," Proc. Workshop Storage Security and Survivability, 2007.

14. E. Shen, E. Shi, and B. Waters, "Predicate Privacy in Encryption Systems" Proc. Sixth Theory of Cryptography Conf. Theory of Cryptography (TCC), 2009.

15. W.K. Wong, D.W. Cheung, B. Kao, and N. Mamoulis, "Secure kNN Computation on Encrypted Databases," Proc. 35th ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 139-152, 2009.

16. D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. of IEEE Symposium on Security and Privacy'00, 2000.

17. Y.-C. Chang and M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," in Proc. of ACNS'05, 2005.

18. R. Curtmola, J. A. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions," in Proc. of ACM CCS'06, 2006.

19. M. Bellare, A. Boldyreva, and A. ONeill, "Deterministic and Efficiently Searchable Encryption," Proc. 27th Ann. Int'l Cryptology Conf. Advances in Cryptology (CRYPTO '07), 2007.

20. C. Wang, N. Cao, K. Ren, and W. Lou, "Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 8, pp. 1467- 1479, Aug. 2012.